Consistent estimation of a general nonparametric regression function in time series

Linton, Oliver; Sancetta, Alessio

Consistent Estimation of A General Nonparametric Regression Function in Time Series

Oliver Linton*
The London School of Economics

Alessio Sancetta†
Cambridge University

21st October 2008

Abstract

We propose an estimator of the conditional distribution of $X_t|X_{t-1}, X_{t-2}, \ldots$, and the corresponding regression function $\mathbb{E}(X_t|X_{t-1}, X_{t-2}, \ldots)$, where the conditioning set is of infinite order. We establish consistency of our estimator under stationarity and ergodicity conditions plus a mild smoothness condition.

Key Words: Kernel; Regression; Time Series

Classification : C12

*Thanks to the ESRC and Leverhulme foundation for financial support. Department of Economics, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom. e-mail: o.linton@lse.ac.uk; web page: http://econ.lse.ac.uk/staff/olinton/

†Thanks to Brendan Beare for a discussion about functions of bounded variations. Faculty of Economics, University of Cambridge, Sidgwick Avenue, Cambridge CB3 9DD, United Kingdom. e-mail: asancetta@gmail.com. web page: http://www.sancetta.googlepages.com

1 Introduction

There are now many papers on nonparametric estimation in time series. Roussas (1967), Rosenblatt (1970,1971) and Pham Dinh Tuan (1981) gave CLT's for kernel density and/or regression function estimators under the Markov hypothesis. Robinson (1983) relaxed the Markov assumption. He studied the case where a sample $\{X_t; t = 1, \ldots, n\}$ is observed where $(X_t)_{t\in\mathbb{Z}}$ is a real-valued stationary and strong mixing stochastic process. The objects of interest were the marginal and conditional density functions as well as the regression function $\mathbb{E}(Y_t|Z_t)$, where Y_t and Z_t are (different) finite dimensional vectors containing lags of X_t. He provided sufficient conditions for the pointwise consistency and asymptotic normality of the kernel estimators under weak dependence. As is by now well known, he found that the rate of convergence and the asymptotic distribution were the same as if the variable X_t was i.i.d. with the same marginal distribution. Robinson (1986) considered also the case of regression where effectively X_t is a vector and Y_t, Z_t are functions of different components of X_t. These results have recently been generalized to local polynomial estimators in Masry and Fan (1997) under more or less the same regularity conditions. Lu and Linton (2007) have extended these results to near epoch dependent functions of mixing processes. Collomb (1985) and Masry (1996) have studied uniform strong convergence. When the assumption of stationarity is abandoned one can find quite different results, for example those obtained by Phillips and Park (1998) and Karlsen and Tjøstheim (2001) for unit root or null recurrent processes (see also Bandi, 2004, for near-integrated processes) for which the rates of convergence are slower and limiting distributions are non-normal (see also Sancetta, 2007b, for modified estimators that lead to standard inference in some of these situations). As remarked in Gyorfi et. al. (1998), while many mixing/dependence conditions seem very plausible, there is virtually no literature on inference for mixing parameters estimated from data.

Hence, following a second strand of literature concerned with consistency only (e.g. Ornstein, 1978, Algoet, 1992, Morvai et. al., 1996) we maintain the hypothesis that the data are stationary, but only require ergodicity. We generalize the object of interest to allow for infinitely many conditioning variables. In particular, we study the estimation of the infinite order regression

$$\mathbb{E}(X_t|\mathcal{F}_{t-1}), \tag{1}$$

where $\mathcal{F}_{t-1} = \sigma(X_s; s < t)$ is the sigma algebra generated by the sequence $(X_s)_{s<t}$. Using the previous notation, we could have $X_t = (Y_t, Z_t)$ so that both the regression and autoregression problems are covered. This object is of interest for the following reasons. First, many parametric time series models involve dependence on the infinite past. For example, the class of linear processes $X_t = \sum_{j=1}^{\infty} c_j(\theta) X_{t-j} + \varepsilon_t$, where $c_j(\theta)$ are coefficients depending on unknown parameters such that $\sum_{j=1}^{\infty} c_j^2(\theta) < \infty$, while ε_t is i.i.d. This class contains the stationary and invertible ARMA processes that are widely used in practice. If X_t is a GARCH(1,1) process, then $X_t^2 = \sum_{j=1}^{\infty} c_j(\theta) X_{t-j}^2 + \varepsilon_t$, although in that case ε_t is not i.i.d. In the Gaussian linear process case the univariate conditional expectations $E(X_t|X_{t-j})$ are also linear and can be used to identify the process. But in the non-Gaussian case this is no longer true and the univariate conditional expectations can be nonlinear and quite different from $\mathbb{E}(X_t|\mathcal{F}_{t-1})$, Tong (1990), thereby making identification of the correct model from just these univariate quantities impossible. Second, there are some semiparametric time series models that are difficult to estimate without some preliminary estimator of $\mathbb{E}(X_t|\mathcal{F}_{t-1})$. To be specific, Linton and Perron (2003) studied the risk premium model

$$X_t = \mu(\sigma_t^2) + \varepsilon_t \sigma_t, \ t = 1, 2, \ldots, n, \tag{2}$$

where ε_t are i.i.d. with zero mean and variance one, σ_t^2 is a GARCH or EGARCH volatility process, while $\mu(\cdot)$ is of unknown functional form. The restriction that $\mathbb{E}(X_t|\mathcal{F}_{t-1})$, where $\mathcal{F}_{t-1} = \sigma(X_s; s < t)$, only depends on the past through σ_t^2 is quite severe but is a consequence of asset pricing models such as for example Backus and Gregory (1992) and Gennotte and Marsh (1988). To estimate the function $\mu(.)$ and the parameters of σ_t^2 they proposed an iterative procedure whose properties have not been established as yet. If one can obtain consistent estimates of $\mu_t = E(X_t|\mathcal{F}_{t-1})$, one can use these as starting values in that algorithm and straightforward arguments can be used to show consistency of the resulting estimates of the function $\mu(.)$ and the parameters of σ_t^2. See also Pagan and Hong (1991) and Pagan and Ullah (1988). A third reason for estimating the unrestricted regression $\mathbb{E}(X_t|\mathcal{F}_{t-1})$ is for specification testing of nonparametric, semiparametric or parametric models. For example, the martingale hypothesis is that $\mathbb{E}(X_t|\mathcal{F}_{t-1}) = 0$ a.s.

The closest work to ours is Morvai et. al. (1996). This paper proposes what amounts to sequential histogram versions of our estimators. Their primary construction is for the

case where the series is binary: they average over the random number of cases where an increasing finite sequence is reproduced. They then generalize to allow for continuous distributions by "quantizing" the sample space, covering it by a partition that refines with sample size. Their estimator involves some implicit temporal downweighting but it is not very transparent because of the sequential nature of its construction. In practice, it is likely to require much greater sample sizes than ours for reasonable performance. Furthermore, it is hard to frame the issues of "quantization" selection. They establish strong consistency of their c.d.f. estimator (in the weak topology of distributions) and regression estimator (under an additional condition of boundedness). Morvai et. al. (1997) propose a modification of this estimator that effectively decouples the quantization from the length of history considered. They show weak consistency results.

Our estimator is relatively simple to implement, and it is intuitively connected with the standard kernel regression estimator, and is very explicit in terms of the spatial and temporal downweighting involved in its construction. Our results provide conditions for uniform strong consistency (existing results deal with the nonuniform case) and are applicable to data in arbitrary metric spaces endowed with a bounded metric and with a partial order ($\leq$). This is of interest when we deal with particular data sequences like functional data (e.g. Ferraty and Vieu, 2007, and Masry, 2005, for results under mixing conditions). An example of such data is when we observe sequences of interest rates term structures and we wish to predict the whole yield curve.

Our theory requires tuning of two parameters and we provide suggestions on how to choose them. One extra condition that we need to impose is some smoothness of the conditional distribution function. This is the price to be paid for using an estimator as simple as the one proposed here and that allows for uniform strong consistency.

In the simplest case, the function $\mathbb{E}(X_t|\mathcal{F}_{t-1})$ is a function from $\mathbb{R}^\infty$ to $\mathbb{R}$, denote it by f. When X_t is weakly dependent, we can expect the influence of lagged values to decay in terms of the modulus of uniform continuity of f,

$$|f(x_i; i \geq 1) - f(z_i; i \geq 1)| \leq \sum_{i \geq 1} a_i |x_i - z_i|^b,$$

where $|\bullet|$ is a suitable norm, $b > 0$ and $a_i \to 0$ as $i \to \infty$. For geometrically mixing X_t we expect that $a_i \sim c\lambda^i$ for some $\lambda \in (0,1)$ and positive constant c. We do not impose such

specific assumptions. Here, we shall only assume that $(X_t)_{t\in\mathbb{Z}}$ is an ergodic and stationary sequence. Additional conditions related to existence of moments and mild smoothness conditions on the conditional distribution function will also be imposed.

2 The Estimator

We assume that we have a backward expanding sample X_{-n}^{-1} of n observations and we are interested in constructing an estimator of $\mathbb{E}\left(X_0|\mathcal{F}_{-1}\right)$. By stationarity and the shift operator, this is equivalent to finding an estimator of $\mathbb{E}\left(X_t|\mathcal{F}_{t-1}\right)$ using X_{t-n}^{t-1}. See Gyorfi et al. (2002, Ch.27) for remarks about estimation using a backward expanding sample and the more challenging estimation based on the forward expanding sample X_1^n.

Our estimator is a locally weighted average, like classical nonparametric regression estimators. The only difference here is the way we must define local, which must take account of the size of the conditioning set. We require some additional details. We let X_t take values in some metric space $(\mathcal{X}, d)$. The product space $\mathcal{X}^\infty = \bigotimes_{s=1}^\infty \mathcal{X}$ is equipped with the metric $\mathbf{d}_\lambda(x,y) = \sum_{s=1}^\infty \lambda^s d(x_s, y_s)$, $x, y \in \mathcal{X}^\infty$, for some $\lambda \in (0,1)$. With abuse of notation, the same $\mathbf{d}_\lambda$ is also used on a finite product space: for $x, y \in \mathcal{X}^n$, $\mathbf{d}_\lambda(x,y) = \sum_{s=1}^n \lambda^s d(x_s, y_s)$. Define the following set of $\mathbf{d}_\lambda$ radius h around x_{-n}^{-1} as

$$B_h\left(x_{-n}^{-1}\right) := \left\{ y \in \mathcal{X}^\infty : \sum_{s=1}^n \lambda^s d\left(x_{-s}, y_s\right) \leq h \right\}. \tag{3}$$

The set $B_h(x)$ includes the set $\tilde{B}_{h/\lambda^s}(x_s) := \{y, x \in \mathcal{X}^\infty : d(y_s, x_s) \leq h/\lambda^s, y_t = x_t, t \neq s\}$, which expands as $s \to \infty$ for fixed h. This means that the neighborhood system has a tilted geometry where distant lags (large s) count much less in the determination of whether a vector is close to another one. Then, for $x \in \mathcal{X}$, we propose the following estimator

$$\mathbb{P}_n\left(x|B_h\left(X_{-n}^{-1}\right)\right) := \frac{\sum_{s=1}^{(n-m)} \left\{X_{-s} \leq x\right\} K\left(\mathbf{d}_\lambda\left(X_{-(n-s)}^{-1}, X_{-n}^{-(1+s)}\right)/h\right)}{\sum_{s=1}^{(n-m)} K\left(\mathbf{d}_\lambda\left(X_{-(n-s)}^{-1}, X_{-n}^{-(1+s)}\right)/h\right)}, \tag{4}$$

where the inequality is meant elementwise if required (e.g. $\mathcal{X} \subseteq \mathbb{R}^v$, $v > 1$), and K is a kernel that has support $[0,1]$. Throughout the paper, for any set A, the indicator function

of the set is written as the set itself: $I_A = A$. The parameter $m \geq 1$ is fixed and chosen such that enough observations are available for reasonable conditional estimation. This can reduce the bias in finite samples. Asymptotically, the value of m is irrelevant, hence, for simplicity we just set it equal to one with no further discussion. The parameter h defines the size of the local conditioning sets and is such that $h := h_n \to 0$ as $n \to \infty$. Finally, $\lambda \in (0,1)$ determines the shape and allocation of the local conditioning sets. These quantities will be implicitly specified in our regularity conditions below.

We define the corresponding estimator of the conditional expectation of some function $g(X_0)$,

$$\begin{aligned} \mathbb{P}_n\left(g(X_0)\,|B_h\left(X_{-n}^{-1}\right)\right) := & \int_{\mathcal{X}} g(x)\,\mathbb{P}_n\left(dx|B_h\left(X_{-n}^{-1}\right)\right) \\ = & \frac{\sum_{s=1}^{n-1} g(X_{-s})\,K\left(\mathbf{d}_\lambda\left(X_{-(n-s)}^{-1}, X_{-n}^{-(1+s)}\right)/h\right)}{\sum_{s=1}^{n-1} K\left(\mathbf{d}_\lambda\left(X_{-(n-s)}^{-1}, X_{-n}^{-(1+s)}\right)/h\right)}. \end{aligned} \tag{5}$$

This can be seen as a form of the Nadaraya-Watson kernel regression estimator with covariates of increasing dimension, but where the influence of temporally remote covariates is small.

3 Main Results

The goal of this section is to state general high level conditions that ensure consistency in a general framework. For simplicity we shall restrict our attention to the uniform kernel case where $K(u) = \{|u| \leq 1\}$. For the estimator in (4) we shall show that

$$\sup_{x\in\mathcal{X}}\left|\mathbb{P}_n\left(x|B_h\left(X_{-n}^{-1}\right)\right) - \Pr\left(X_0 \leq x|\mathcal{F}_{-1}\right)\right| \to 0,$$

in probability (a.s.). For the estimator in (5), in the special case $g(x) = x^p$, $p \in \mathbb{N}$, the previous display together with an additional regularity condition implies

$$\mathbb{P}_n\left(X_0^p|B_h\left(X_{-n}^{-1}\right)\right) := \int_{\mathcal{X}} x^p \mathbb{P}_n\left(dx|B_h\left(X_{-n}^{-1}\right)\right) \to \mathbb{E}\left(X_0^p|\mathcal{F}_{-1}\right),$$

in probability (a.s.).

We formally state the conditions that imply consistency of the estimator.

Condition 1 $(X_t)_{t\in\mathbb{Z}}$ *is a stationary and ergodic sequence of random variables with law* $\mathbb{P}$ *and values in* $\mathcal{X}$ *endowed with a partial order* $\leq$.

We impose smoothness on the joint distribution function.

Condition 2 *The conditional probability* $\Pr\left(X_0 \leq \bullet | X_{-\infty}^{-1} = x_{-\infty}^{-1}\right)$ *is* $\mathbb{P}$ *a.s. continuous in* $x_{-\infty}^{-1}$ *with respect to the topology generated by* $\mathbf{d}_\lambda$.

The next is the crucial condition for consistency.

Condition 3 *For* $\mathbb{P}$*-almost all* x_{-n}^{-1}*, choose* $h_n \to 0$ *such that*

$$\lim_{n\to\infty} \sum_{s=1}^{n-1} \left\{ \mathbf{d}_\lambda \left(x_{-(n-s)}^{-1}, X_{-n}^{-(1+s)} \right) \leq h_n \right\} = \infty \text{ in probability } (a.s.).$$

In Section 4.1 we provide a simple condition on the metric d that is sufficient for Condition 3 to be non-vacuous. Hence, we have the following.

Theorem 1 *Suppose that the family of sets* $\{\{s \in \mathcal{X} : s \leq x\}; x \in \mathcal{X}\}$ *has finite bracketing number. Under Conditions 1, 2 and 3,*

$$\sup_{x\in\mathcal{X}} \left| \mathbb{P}_n \left(x | B_h \left(X_{-n}^{-1} \right) \right) - \Pr\left(X_0 \leq x | \mathcal{F}_{-1} \right) \right| \to 0 \text{ in probability } (a.s.).$$

If $\mathcal{X} \subseteq \mathbb{R}^v$, the left open intervals $\{\{s \in \mathcal{X} : s \leq x\}; x \in \mathcal{X}\}$ have finite bracketing numbers for v bounded (van der Vaart and Wellner, 2000). We can use a uniform integrability condition to show a related uniform convergence result for classes of functions which we denote by $\mathfrak{G}$.

Condition 4 $\mathfrak{G}$ *is a family of functions with envelope function* $G(x) = \sup_{g\in\mathfrak{G}} |g(x)|$ *such that*

$$\sup_{1\leq i\leq n<\infty} \mathbb{E}\left[G\left(X_{-i+1}\right)^p \mid \left\{ \mathbf{d}_\lambda \left(X_{-(n-i+1)}^{-1}, X_{-n}^{-i} \right) \leq h_n \right\} \right] < \infty, \text{ for some } p > 1.$$

Condition 4 makes sure that the terms in the summation defining (5) are uniformly integrable. (Note that summation is over the X_i's satisfying $\{\mathbf{d}_\lambda(X_{-(n-i+1)}^{-1}, X_{-n}^{-i}) \leq h_n\}$). We now state two corollaries to Theorem 1 that follow by use of Condition 4.

Corollary 1 *Let $\mathfrak{G}$ be the family of equicontinuous functions satisfying Condition 4. Then, under Conditions 1, 2 and 3,*

$$\sup_{g\in\mathfrak{G}}\left|\int_{\mathcal{X}} g(x)\,\mathbb{P}_n\left(dx|B_h\left(X_{-n}^{-1}\right)\right) - \mathbb{E}\left(g\left(X_0\right)|\mathcal{F}_{-1}\right)\right| \to 0 \text{ in probability } (a.s.).$$

For example, a family of functions is equicontinuous if it contains functions that are Lipschitz under some metric or if it comprises of a finite arbitrary collection of continuous functions.

Remark 1 *Clearly, when $\mathfrak{G}$ comprises of the single function x^p, $p \in \mathbb{N}$, which is continuous, Corollary 1 implies consistency for conditional moment estimators, i.e. $\mathbb{E}\left(X_0^p|\mathcal{F}_{-1}\right)$.*

In some circumstances, we are interested in $\mathfrak{G}$ whose elements are not necessarily continuous. If we restrict attention to functions with domain in a Euclidean space $\mathcal{X} \subseteq \mathbb{R}^v$ (v a finite integer), the elements in $\mathfrak{G}$ can be replaced by functions of Hardy bounded variation. We briefly recall the definition before stating the result.

Definition 1 *A function $g : \mathbb{R}^v \to \mathbb{R}$ is of Hardy bounded variation (BV) if it can be written as $g(x) = g_1(x) - g_2(x)$ where g_j $(j = 1, 2)$ are coordinatewise increasing functions, finite on any compact subset of $\mathcal{X}$.*

Note that for $v = 1$ all definitions of bounded variation are the same and they differ for $v > 1$ (e.g. Clarkson and Adams, 1933). Hence, we have the following.

Corollary 2 *Suppose that $\mathfrak{G}$ is a class of BV functions satisfying Condition 4. Then, under Conditions 1, 2 and 3,*

$$\sup_{g\in\mathfrak{G}}\left|\int_{\mathcal{X}} g(x)\,\mathbb{P}_n\left(dx|B_h\left(X_{-n}^{-1}\right)\right) - \mathbb{E}\left(g\left(X_0\right)|\mathcal{F}_{-1}\right)\right| \to 0 \text{ in probability } (a.s.).$$

Note that continuous functions are not necessarily BV function, e.g. $g(x) = x\sin(1/x)$ for $x > 0$, and zero elsewhere, is continuous, but not of bounded variation. Basically, functions of Hardy bounded variation are functions having a.e. the derivative $\mathcal{D}^v g$, where $(\mathcal{D}^v g)(x) = \partial^v g(x) / (\partial x_1 \cdots \partial x_v)$, $x = (x_1, \ldots, x_v)$. We now turn to some further discussion.

4 Discussion

4.1 Remarks on Condition 3

While we do not impose dependence conditions, verification of Condition 3 is a major difficulty, but it is exactly what is required for consistency. If Condition 3 holds, there is no need to require the data sequence to be ergodic. Nevertheless, ergodicity appears to be needed in order to verify Condition 3. Recall that Condition 3 relates to the way the bandwidth needs to be chosen. Condition 1 does not seem to imply that there exists a bandwidth for which Condition 3 holds. For stationary ergodic processes, recurrence to some set is implied by the Poincare Recurrence Theorem (e.g. Theorem 6.4.1 in Gray 1998). In our case, the set is expanding and we cannot make direct use of this result. However, under an additional mild technical condition we can show that Condition 1 is sufficient to ensure that Condition 3 can be satisfied.

Condition 5 *The metric d is bounded, i.e. $\max_{x,y\in\mathcal{X}} d(x,y) \leq C$, where C is a finite absolute constant.*

This condition has minor practical consequences. Indeed we can easily turn any metric d' on $\mathcal{X}$ into a bounded one, e.g. $d := d'/(1+d')$. Then, we have the following.

Lemma 1 *Under Conditions 1 and 5, there is a sequence $h_n \to 0$ such that, for $\mathbb{P}$ almost all $x_{-\infty}^{-1}$,*

$$\lim_{n\to\infty} \sum_{s=1}^{n} \left\{ \mathbf{d}_\lambda \left(x_{-(n-s)}^{-1}, X_{-n}^{-(1+s)} \right) \leq h_n \right\} = \infty \, a.s.$$

In the proof of Lemma 1, it is shown that

$$B_h^a\left(x_{-I}^{-1}\right) := \left\{ d\left(x_{-i}, X_{-(s+i)}\right) \leq a_i \frac{h}{\lambda^i}; i = 1, \ldots, I \right\} \subset \left\{ \mathbf{d}_\lambda \left(x_{-(n-s)}^{-1}, X_{-n}^{-(1-s)} \right) \leq h \right\}$$

for any sequence $a := (a_i)_{i>0}$ such that $a_i \geq a_{i+1}$, $\sum_{i>0} a_i \leq 1$ and

$$I = \inf \left\{ i \in \mathbb{N} : \left(a_i h / \lambda^i \right) \geq 1 \right\}.$$

Clearly, I depends on h and a. Hence, to check Condition 3 we can check that

$$\sum_{s=1}^{n-I} \left\{ X_{-(s+I)}^{-(s+1)} \in B_h^a \left(x_{-I}^{-1} \right) \right\} \to \infty \tag{6}$$

in probability (a.s.). Then, (6) is similar in spirit to standard conditions used to show convergence of kernel regression estimators (e.g. Devroye, 1981, Theorem 4.1). It would be conceptually useful to relate $h_n \to 0$ directly to n. Suppose that for $\mathbb{P}$ almost all x_{-I}^{-1},

$$\frac{1}{R_n} \left| \sum_{s=1}^{n} \left\{ X_{-(s+I)}^{-(s+1)} \in B_h^a \left(x_{-I}^{-1} \right) \right\} - \sum_{s=1}^{n} \Pr \left(X_{-(s+I)}^{-(s+1)} \in B_h^a \left(x_{-I}^{-1} \right) \right) \right| \to 0 \tag{7}$$

in probability (a.s.) for some sequence $R_n = R_n(h) = o\left(n \Pr \left(X_{-(1+I)}^{-(1+1)} \in B_h^a \left(x_{-I}^{-1} \right) \right) \right)$. By recurrence, the sequence $R_n \to \infty$ only when $I = o(n)$ implying (6), hence Condition (3). To show (7) we would need regularity conditions on $\Pr \left(X_{-(1+I)}^{-(1+1)} \in B_h^a \left(x_{-I}^{-1} \right) \right)$ in order to find its rate of decay as well as suitable mixing conditions (e.g. Rio, 2000, for a review). Given that x_{-I}^{-1} expands as $n \to \infty$ the resulting conditions on h_n are very complex and can be only stated as the solution of some nonlinear equation. Hence, for the sake of simplicity (and generality) our results are presented under Condition 3 only without using mixing conditions. Nevertheless, having established that Condition 3 is not void, it is necessary to choose h_n in some reasonable way. We discuss this issue next.

4.2 Remarks on Parameter Selection

Estimators (4) and (5) depend on parameters $h \to 0$ and $\lambda \in (0, 1)$ and it is not obvious a priori what are good choices of them. The weak conditions used here make the direct application of classical cross-validation procedures difficult and possibly dubious. In fact, while cross-validation for time series has been considered in the literature (Härdle and Vieu, 1992), the conditions required are too strict for the present context. In particular, the proof for the consistency of crossvalidation in Härdle and Vieu (1992) relies on inequalities

for moments of partial sums (i.e. Marcinkiewicz–Zygmund kind of inequalities; e.g. see their Lemmata 3 and 4). Related moment inequalities are also used to derive the rate of convergence of the nonparametric estimator to the true regression function (their Lemma 1). None of these results is applicable here. Hence, we are only left with the choice of splitting the sample into an estimation sample and a validation sample over which to evaluate the performance of different bandwidths. Clearly, the splitting could be done recursively leading to a procedure that is amenable to standard analysis. For the sake of clarity we outline the procedure. Let $\mathbb{P}_n\left(\bullet|B_h\left(X_1^n\right)\right)$ be (4) where we have shifted forward the segment of observations $(X_{-1},\ldots,X_{-n})$ used to construct the estimator. Parametrize the possible sequence of smoothing parameters, i.e. $h = h_n(\beta)$. Then, the problem reduces to optimal choice of $\pi := (\lambda,\beta)$ with $\pi \in \Pi \subset \mathbb{R}^2$. The problem reduces to forecast validation as done in the prequential statistical literature (Dawid, 1997, for a review and references). The estimators discussed in this paper are functions of $\mathbb{P}_n(\bullet|\pi) = \mathbb{P}_n\left(\bullet|B_h\left(X_1^n\right)\right)$ (emphasizing dependence on π). We only discuss the regression problem $\mathbb{P}_n\left(X_{n+1}|\pi\right) = \int_{\mathcal{X}} x\mathbb{P}_n\left(dx|\pi\right)$. Let $\mathbb{E}_n$ be expectation conditioning on $\mathcal{F}_n$. Define

$$\bar{\mathcal{L}}_N(\theta) = \sum_{n=m}^{N} \mathbb{E}_n\left|X_{(n+1)} - \mathbb{P}_n\left(X_{(n+1)}|\pi\right)\right|^2$$

so that minimization of $\bar{\mathcal{L}}_N(\pi)$ with respect to π delivers the forecast closest to the conditional mean, say $\pi_{(N)}$. Since $\bar{\mathcal{L}}_N(\pi)$ is unknown, we minimize the empirical criterion

$$\mathcal{L}_N(\pi) := \sum_{n=m}^{N} \left|X_{(n+1)} - \mathbb{P}_n\left(X_{(n+1)}|\pi\right)\right|^2$$

$$\hat{\pi}_{(N)} := \arg\min_{\pi\in\Pi} \mathcal{L}_N(\pi).$$

By the martingale structure of $\mathcal{L}_N(\pi) - \bar{\mathcal{L}}_N(\pi)$, under regularity conditions, the empirical optimal choice $\hat{\pi}_{(N)}$ can be shown to be close to $\pi_{(N)}$ in probability, using standard martingale arguments (e.g. Seillier-Moiseiwitsch and Dawid, 1993).

5 Numerical Work

5.1 Simulation

In this section we discuss some Monte Carlo results whose aim is to verify the consistency of a simple implementation of our procedure. We suppose that

$$X_t = 1 + \varepsilon_t - \theta\varepsilon_{t-1},$$

where ε_t is either $N(0, \sigma^2)$ or $U[-\sigma/2, \sigma/2]$. When ε_t is Gaussian, the conditional expectation $\mathbb{E}(X_t|X_{t-1}) = 1 - \theta(X_{t-1} - 1)/(1 + \theta^2)$ is linear, but when ε_t is uniform, it can be nonlinear, see Tong (1990, pp 13-14). But in either case, $\mathbb{E}(X_t|X_{t-1}, \ldots) = 1 - \theta\varepsilon_{t-1} = 1 - \theta(X_{t-1} - 1)/(1 - \theta L)$ (L is the lag operator), which depends linearly on all past values of X. This is assuming invertibility, i.e., $|\theta| < 1$.

We consider a fixed sample size $n = 1000$ and change the parameter $\sigma \in \{0.01, 0.1, 0.3, 1\}$. The effect of decreasing error scale should be similar to that of increasing sample size. We consider $\theta \in \{0.0, 0.33, 0.66, 1.0\}$.

We have used $d(x, y) = |x - y|$. We set $\lambda = \widehat{\lambda} = \sum_t (X_t - \overline{X})(X_{t-1} - \overline{X}) / \sum_t (X_{t-1} - \overline{X})^2$. This seems to capture the idea that the more dependent X_t is, the larger we should set λ. We have tried other, fixed, values of λ and found similar results. To choose the value of h we have just taken h such that two hundred neighbors are included. Let $g = \mathbb{P}\left(X_0|B_h\left(X_{-n}^{-1}\right)\right)$ and $\widehat{g} = \mathbb{P}_n\left(X_0|B_h\left(X_{-n}^{-1}\right)\right)$, and define also the one dimensional estimators $\widehat{g}_1 = \mathbb{P}_n\left(X_0|B_h\left(X_{-1}^{-1}\right)\right)$.

In Table 1 below we report the bias $\mathbb{E}\widehat{g} - g$ and standard deviation $std(\widehat{g})$ for the uniform error case, where both moments are computed by averaging across the one thousand simulations. The results improve as σ decreases and as θ decreases, but even when $\theta = 1$, the estimator appears consistent. Note that $\widehat{g}_1$ is inconsistent in this case. The results for the normal distribution are similar and not shown.

Table 1

σ/θ	0.0		0.33		0.66		1.0	
	bias	std	bias	std	bias	std	bias	std
1.0	0.0055	0.2732	0.0008	0.2884	-0.0081	0.3400	0.0060	0.4226
0.3	-0.0027	0.0823	-0.0008	0.0855	-0.0028	0.1005	-0.0038	0.1336
0.1	0.0001	0.0273	0.0012	0.0278	0.0002	0.0329	-0.0016	0.0439
0.01	0.0001	0.0027	0.0001	0.0028	0.0000	0.0034	0.0002	0.0043

The distribution of the estimator appears approximately normal according to Figure 1.

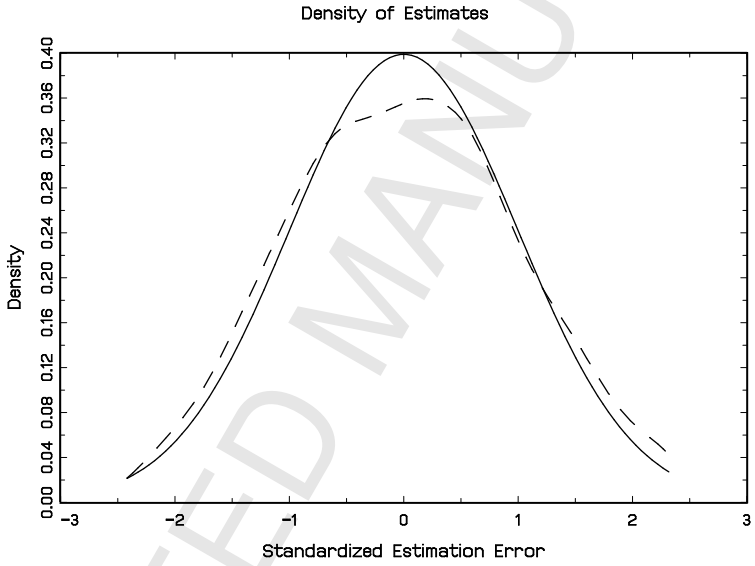

Figure 1. This shows the case where $\theta = 1$ and $\sigma = 1$. Solid line is the standard normal pdf, dashed line is the estimated density of $\widehat{g} - g$ (standardized to have mean zero and variance one).

In Table 2 we show the case where $\sigma = 0.3$ for different sample sizes $n \in \{100, 400, 1600, 6400\}$. This shows that consistency (as $n \to \infty$) is achieved but the convergence is rather slow.

Table 2

n/θ	0.0		0.33		0.66		1.0	
	bias	std	bias	std	bias	std	bias	std
100	0.0000	0.0120	-0.0009	0.0173	0.0006	0.0370	0.0032	0.0697
400	-0.0001	0.0089	0.0000	0.0126	0.0000	0.0337	0.0001	0.0629
1600	0.0001	0.0061	0.0007	0.0107	0.0000	0.0326	-0.0065	0.0627
6400	0.0000	0.0043	0.0002	0.0094	-0.0004	0.0309	-0.0013	0.0615

5.2 Application

We apply our theory to the study of the risk return relationship. Modern asset pricing theories imply restrictions on the time series properties of expected returns and conditional variances of market aggregates. These restrictions are generally quite complicated, depending on utility functions as well as on the driving process of the stochastic components of the model. However, in an influential paper, Merton (1973) obtained very simple restrictions albeit under somewhat drastic assumptions; he showed in the context of a continuous time partial equilibrium model that

$$\mu_t = E[(r_{mt} - r_{ft})|\mathcal{F}_{t-1}] = \gamma \text{var}[(r_{mt} - r_{ft})|\mathcal{F}_{t-1}] = \gamma\sigma_t^2, \tag{8}$$

where r_{mt}, r_{ft} are the returns on the market portfolio and risk-free asset respectively, while $\mathcal{F}_{t-1}$ is the market wide information available at time $t-1$. The constant γ is the Arrow–Pratt measure of relative risk aversion. The linear functional form actually only holds when σ_t^2 is constant; otherwise μ_t and σ_t^2 can be nonlinearly related, Gennotte and Marsh (1993). Many previous tests of this restriction imposed parametric specifications both in the dynamics of the volatility process σ_t^2 like GARCH-M and in the relationship between risk and return like linearity. Pagan and Hong (1990) argue that the risk premium μ_t and the conditional variance σ_t^2 are highly nonlinear functions of the past whose form is not captured by standard parametric GARCH–M models. They estimate μ_t and σ_t^2 as nonparametric regressions on a finite dimensional information set finding evidence of considerable nonlinearity. They then estimated γ from the regression $r_{mt} - r_{ft} = \gamma\sigma_t^2 + \eta_t$, by least squares and instrumental variables methods with σ_t^2 substituted by the nonparametric estimate, finding a negative but insignificant γ. Linton and Perron (2003) considered

the model (2), where σ_t^2 was a parametrically specified CH process (with dependence on the infinite past) but $\mu_t = \varphi(\sigma_t^2)$ for some function φ of unknown functional form. They proposed an estimation algorithm but did not establish any statistical properties. They found some evidence of a nonlinear relationship.

We suppose that both functions μ_t and σ_t^2 are unrestricted nonparametric functions of the entire information set $\mathcal{F}_{t-1}$ and they are related in a general way, that is, $\mu_t = \varphi(\sigma_t^2)$ for some function φ of unknown functional form, or equivalently $X_t = \varphi(\sigma_t^2) + \eta_t$, where η_t is a martingale difference sequence satisfying $\mathbb{E}(\eta_t|\mathcal{F}_{t-1}) = 0$. Below we show some preliminary estimation of $\mu_t = \mathbb{E}(X_t|\mathcal{F}_{t-1})$ and $\sigma_t^2 = \mathrm{var}(X_t|\mathcal{F}_{t-1})$ using S&P500 weekly stock returns with $n = 2475$. We chose $\lambda = 0.99$ and h such that $k = 200$ lags were included in the weighting. We then estimated the function φ by a univariate local linear kernel estimator with Silverman rule of thumb bandwidth. We show the estimated function φ. The relationship is rather weak, i.e., the function φ is close to a constant.

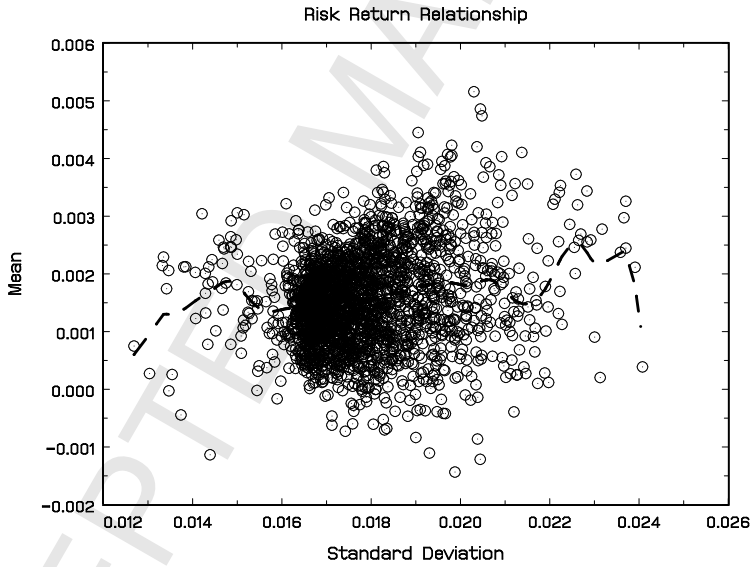

Figure 2. Weekly returns on the S&P500 Index. On the horizontal axis is shown the square root of the estimated conditional variance of returns given the infinite past; on the vertical axis is shown the estimated conditional mean of returns given the infinite past. The dashed line is the one-dimensional smooth of estimated mean on estimated standard deviation.

6 Concluding Remarks

We have established the uniform strong consistency of the conditional distribution function estimator under very weak conditions. It is reasonable to expect that the best rate we can hope for over this very large class of functions is logarithmic in sample size. To formally derive almost sure rates of convergence we would need to impose dependence conditions that allow us to control the bias of the estimator. These dependence conditions would also be needed to establish some exponential inequality to control the estimation error. Exponential inequalities are commonly used in the application of the Borel-Cantelli lemma to ensure that the convergence is almost sure. If rates of convergence were available, then we could also hope to derive a central limit theorem for the estimator.

It is an open question whether one can achieve algebraic rates for some restricted class of functions. For example, suppose that

$$f\left(x_i; i \geq 1\right) = \sum_{i=1}^{\infty} f_i(x_i), \tag{9}$$

where the functions $f_i(.)$ are such that the sum is well defined, which implies some decay in their respective magnitudes. This additive regression model has been well studied in the case where it is known that $f_i(.) \equiv 0$ for all $i > d$ for some finite d. Stone (1985) showed that the optimal rate for estimation of the components $f_i(.)$ and $f(.)$ is the same as for one-dimensional nonparametric regression. Estimation algorithms have been proposed in Linton and Nielsen (1995) and Mammen, Linton, and Nielsen (1999). Linton and Mammen (2005) have considered the case where $d = \infty$ but $f_i(x_i) = \psi_i(\theta)m(x_i)$ for some parametric quantities $\psi_i(\theta)$ that decline suitably fast. It may be possible to adapt the algorithm of Mammen, Linton, and Nielsen (1999) to the general model (9) by allowing the number of dimensions iterated over to increase slowly with sample size but such analysis is beyond the scope of this paper.

A Appendix: Proof of Main Results

At first we note the following, simple result.

Lemma 2 *For any* $x_{-\infty}^{-1} \in \mathcal{X}^{\infty}$,

$$\lim_{h\to 0} B_h\left(x_{-\infty}^{-1}\right) = \lim_{h\to 0}\left\{y_{-\infty}^{-1} \in \mathcal{X}^{\infty} : \mathbf{d}_{\lambda}\left(x_{-\infty}^{-1}, y_{-\infty}^{-1}\right) \leq h\right\} = \left\{x_{-\infty}^{-1}\right\}.$$

Proof. For any sequence $k_n \to \infty$

$$\left\{x_{-\infty}^{-1}\right\} \subseteq B_h\left(x_{-\infty}^{-1}\right) \subseteq B_h\left(x_{-k_n}^{-1}\right).$$

Hence it is sufficient to show that

$$\lim_{n\to\infty} B_{h_n}\left(x_{-k_n}^{-1}\right) = \left\{x_{-\infty}^{-1}\right\}.$$

Since

$$B_h\left(x_{-k}^{-1}\right) \subseteq \bigcap_{s=1}^{k}\left\{y \in \mathcal{X} : d\left(x_{-s}, y_{-s}\right) \leq \frac{h}{\lambda^s}\right\}$$

we choose $k = o\left(\log_{1/\lambda}\left(1/h\right)\right)$ so that, as $h \to \infty$, $h/\lambda^k \to 0$ implying

$$\bigcap_{s=1}^{k}\left\{y \in \mathcal{X} : d\left(x_{-s}, y_{-s}\right) \leq \frac{h}{\lambda^s}\right\} \downarrow \bigcap_{s=1}^{\infty}\left\{y_{-s} \in \mathcal{X} : d\left(x_{-s}, y_{-s}\right) = 0\right\} = \left\{x_{-\infty}^{-1}\right\}$$

and the result is proved. ■

Proof. [Theorem 1] Define

$$\tau_{1,n} := \inf\left\{s > 1 : \mathbf{d}_{\lambda}\left(x_{-(n-s-1)}^{-1}, X_{-n}^{-s}\right) \leq h_n\right\}$$

and, for $i \geq 1$,

$$\tau_{i+1,n} := \tau_{i,n} + \inf\left\{s > 0 : \mathbf{d}_{\lambda}\left(x_{-(n+1-\tau_{i,n}-s)}^{-1}, X_{-n}^{-(\tau_{i,n}+s)}\right) \leq h_n\right\}$$

and furthermore

$$I_n := \sup\left\{i \geq 1 : \tau_{i,n} \leq n\right\}. \tag{10}$$

With this notation write, for $\mathbb{P}$-almost all x_{-n}^{-1},

$$\begin{aligned}\mathbb{P}_n\left(x|B_h\left(x_{-n}^{-1}\right)\right) &:= \frac{\sum_{s=1}^{n}\left\{X_{-s}\leq x\right\}\left\{\mathbf{d}_\lambda\left(x_{-(n-s)}^{-1},X_{-n}^{-(1+s)}\right)\leq h\right\}}{\sum_{s=1}^{n}\left\{\mathbf{d}_\lambda\left(x_{-(n-s)}^{-1},X_{-n}^{-(1+s)}\right)\leq h\right\}} \\ &= \frac{\sum_{i=1}^{\infty}\left\{X_{-\tau_{i,n}+1}\leq x\right\}\cap\left\{\tau_{i,n}\leq n\right\}}{\sum_{i=1}^{\infty}\left\{\tau_{i,n}\leq n\right\}} \\ &= \frac{1}{I_n}\sum_{i=1}^{I_n}\left\{X_{-\tau_{i,n}+1}\leq x\right\}.\end{aligned}$$

Hence,

$$\begin{aligned}&\frac{1}{I_n}\sum_{i=1}^{I_n}\left[\left\{X_{-\tau_{i,n}+1}\leq x\right\}-\Pr\left(X_0\leq x|X_{-\infty}^{-1}=x_{-\infty}^{-1}\right)\right] \\ =\ &\frac{1}{I_n}\sum_{i=1}^{I_n}\left[\left(1-\mathbb{E}_i\right)\left\{X_{-\tau_{i,n}+1}\leq x\right\}\right] \\ &+\frac{1}{I_n}\sum_{i=1}^{I_n}\left[\mathbb{E}_i\left\{X_{-\tau_{i,n}+1}\leq x\right\}-\Pr\left(X_0\leq x|X_{-\infty}^{-1}=x_{-\infty}^{-1}\right)\right] \\ =\ &\mathrm{I}+\mathrm{II}\end{aligned}$$

where $\mathbb{E}_i$ is expectation conditioning on $\mathcal{F}_{-\tau_{i,n}}$. Since, by Condition 3, $I_n\to\infty$ in probability (a.s.), $|\mathrm{I}|\to 0$ in probability (a.s.) by the martingale strong law of large numbers. Note that $1\leq\tau_{i,n}<\tau_{i+1,n}\leq n$ for $i=1,...,I_n\leq n$ and $I_n\to\infty$. Hence, for any sequence $J_n\to\infty$ such that $J_n=o\left(I_n\right)$ and $i\leq I_n-J_n$ we must have $(n-\tau_{i,n})\to\infty$ in probability

(a.s.). Moreover, note

$$\begin{aligned}
\text{II} &= \frac{1}{I_n}\sum_{i=1}^{I_n}\left[\mathbb{E}_i\left\{X_{-\tau_{i,n}+1}\leq x\right\}-\Pr\left(X_0\leq x|X_{-\infty}^{-1}=x_{-\infty}^{-1}\right)\right] \\
&= \frac{1}{I_n}\sum_{i=1}^{I_n}\left[\Pr\left(X_{-\tau_{i,n}+1}\leq x|X_{-\infty}^{-\tau_{i,n}}\right)-\Pr\left(X_0\leq x|X_{-\infty}^{-1}=x_{-\infty}^{-1}\right)\right] \\
&= \frac{1}{I_n}\left(\sum_{i=1}^{I_n-J_n}+\sum_{i=(I_n-J_n+1)}^{I_n}\right)\left[\Pr\left(X_{-\tau_{i,n}+1}\leq x|X_{-\infty}^{-\tau_{i,n}}\right)-\Pr\left(X_0\leq x|X_{-\infty}^{-1}=x_{-\infty}^{-1}\right)\right] \\
&= \frac{1}{I_n}\sum_{i=1}^{I_n-J_n}\left[\Pr\left(X_{-\tau_{i,n}+1}\leq x|X_{-\infty}^{-\tau_{i,n}}\right)-\Pr\left(X_0\leq x|X_{-\infty}^{-1}=x_{-\infty}^{-1}\right)\right]+o\left(1\right) \qquad (11)
\end{aligned}$$

because the second sum is $O\left(J_n/I_n\right)=o\left(1\right)$. By definition, $X_{-\infty}^{-\tau_{i,n}}\in B_h\left(x_{-(n+1-\tau_{i,n})}^{-1}\right)$, so that we can explicitly write

$$\Pr\left(X_{-\tau_{i,n}+1}\leq x|X_{-\infty}^{-\tau_{i,n}}\right)=\Pr\left(X_{-\tau_{i,n}+1}\leq x|X_{-\infty}^{-\tau_{i,n}}=y_{-\infty}^{-\tau_{i,n}}\in B_h\left(x_{-(n+1-\tau_{i,n})}^{-1}\right)\right). \qquad (12)$$

Let T be the left shift operator, i.e. $TX_s=X_{s+1}$, $T^kX_s=X_{s+k}$. Then, for any $i\leq I_n-J_n$, using the explicit notation in (12),

$$\begin{aligned}
\lim_n\Pr\left(X_{-\tau_{i,n}+1}\leq x|X_{-\infty}^{-\tau_{i,n}}\right) &= \lim_n\Pr\left(T^{(\tau_{i,n}-1)}X_{-\tau_{i,n}+1}\leq x|T^{(\tau_{i,n}-1)}X_{-\infty}^{-\tau_{i,n}}\right) \\
&\quad \text{[by stationarity using the shift operator } T\text{]} \\
&= \Pr\left(X_0\leq x|X_{-\infty}^{-1}=y_{-\infty}^{-1}\in B_h\left(x_{-\infty}^{-1}\right)\right), \qquad (13)
\end{aligned}$$

because for any $i\leq I_n-J_n$, $(n-\tau_{i,n})\to\infty$ implying that for any $h>0$

$$B_h\left(x_{-(n+1-\tau_{i,n})}^{-1}\right)\to B_h\left(x_{-\infty}^{-1}\right).$$

Since h is arbitrary we can choose $h=h_n\to 0$ as in Condition 3 so that, $B_h\left(x_{-\infty}^{-1}\right)\to\left\{x_{-\infty}^{-1}\right\}$ by Lemma 2. Hence, by Condition 2, the last remark together with (13) implies that for $\mathbb{P}$ almost all $x_{-\infty}^{-1}$,

$$\left|\Pr\left(X_{-\tau_{i,n}+1}\leq x|X_{-\infty}^{-\tau_{i,n}}\right)-\Pr\left(X_0\leq x|X_{-\infty}^{-1}=x_{-\infty}^{-1}\right)\right|\to 0,$$

in probability (a.s.) for all $i \leq I_n - J_n$, implying $|\text{II}| \to 0$ (in 11) in the same mode of convergence because $(I_n - J_n)/I_n \to 1$. Since the result holds for $\mathbb{P}$-almost all $x_{-\infty}^{-1}$, it holds for $X_{-\infty}^{-1}$ as well. Using a finite number of brackets for $\{\{s \in \mathcal{X} : s \leq x\}; x \in \mathcal{X}\}$ the convergence is also uniform in $x \in \mathcal{X}$ (e.g. see the proof of Theorem 2.4.1 in van der Vaart and Wellner, 2000). ■

To prove the corollaries we use the following.

Lemma 3 *Condition 4 implies*

$$\lim_{M\to\infty} \lim_{N\to\infty} \sup_{n>N} \int_{\{x\in\mathcal{X}:G(x)>M\}} G(x)\, \mathbb{P}_n\left(dx|B_h\left(X_{-n}^{-1}\right)\right) = 0 \ a.s. \tag{14}$$

Proof. [Lemma 3] It is well known that a moment condition implies uniform integrability (e.g. Example 1.11.4 in van der Vaart and Wellner, 2000), i.e.

$$\lim_{N\to\infty} \sup_{n>N} \int_{\mathcal{X}} G(x)^p\, \mathbb{P}_n\left(dx|B_h\left(X_{-n}^{-1}\right)\right) < \infty \ a.s.$$

for some $p > 1$ implies (14). Define

$$\tau_{1,n} := \inf\left\{s > 1 : \mathbf{d}_\lambda\left(X_{-(n-s-1)}^{-1}, X_{-n}^{-s}\right) \leq h_n\right\}$$

and, for $i \geq 1$,

$$\tau_{i+1,n} := \tau_{i,n} + \inf\left\{s > 0 : \mathbf{d}_\lambda\left(X_{-(n+1-\tau_{i,n}-s)}^{-1}, X_{-n}^{-(\tau_{i,n}+s)}\right) \leq h_n\right\}$$

which is just the sequence of stopping times defined in the proof of Theorem using $X_{-(n+1-\tau_{i,n}-s)}^{-1}$ instead of $x_{-(n+1-\tau_{i,n}-s)}^{-1}$. Hence, mutatis mutandis, define I_n as in (10). Rewrite

$$\int_{\mathcal{X}} G(x)^p\, \mathbb{P}_n\left(dx|B_h\left(X_{-n}^{-1}\right)\right) = \frac{1}{I_n}\sum_{i=1}^{I_n} G\left(X_{-\tau_{i,n}+1}\right)^p.$$

Then, for any $I_n \geq 0$, the above display is a.s. finite if $\sup_{1\leq i\leq n<\infty} \mathbb{E}G\left(X_{-\tau_{i,n}+1}\right)^p < \infty$ for some $p > 1$. By stationarity this is just equal to

$$\begin{aligned}\sup_{1\leq i\leq n<\infty} \mathbb{E}G\left(X_{-\tau_{i,n}+1}\right)^p &= \sup_{1\leq i\leq n<\infty} \mathbb{E}\left[G\left(X_{-\tau_{i,n}+1}\right)^p \mid \left\{\mathbf{d}_\lambda\left(X_{-(n-\tau_{i,n}+1)}^{-1}, X_{-n}^{-\tau_{i,n}}\right) \leq h_n\right\}\right] \\ &\leq \sup_{1\leq i\leq n<\infty} \mathbb{E}\left[G\left(X_{-i+1}\right)^p \mid \left\{\mathbf{d}_\lambda\left(X_{-(n-i+1)}^{-1}, X_{-n}^{-i}\right) \leq h_n\right\}\right] < \infty,\end{aligned}$$

taking the supremum over all $i \leq n$ rather than $\tau_{i,n} \leq n$ only. ■

Proof. [Corollary 1] By Lemma 3 we directly work with (14). Write $\mathbb{P}\left(x|\mathcal{F}_{-1}\right) = \Pr\left(X_0 \leq x|\mathcal{F}_{-1}\right)$ and define $G^M\left(x\right) := G\left(x\right) \wedge M$. For any finite M,

$$\begin{aligned}\lim_{N\to\infty} \sup_{n\geq N} \int_{\mathcal{X}} G\left(x\right) \mathbb{P}_n\left(dx|B_h\left(X_{-n}^{-1}\right)\right) &\geq \lim_{n\to\infty} \int_{\mathcal{X}} G\left(x\right) \mathbb{P}_n\left(dx|B_h\left(X_{-n}^{-1}\right)\right) \\ &\geq \lim_{n\to\infty} \int_{\mathcal{X}} G^M\left(x\right) \mathbb{P}_n\left(dx|B_h\left(X_{-n}^{-1}\right)\right) \\ &= \int_{\mathcal{X}} G^M\left(x\right) \mathbb{P}\left(dx|\mathcal{F}_{-1}\right) \ a.s.\end{aligned}$$

where the equality follows by weak convergence (Theorem 1) because $G^M\left(x\right)$ is continuous and bounded. By asymptotic uniform integrability the left hand side of the above display is finite. Hence, by the monotone convergence theorem

$$\int_{\mathcal{X}} \left[G\left(x\right) - G^M\left(x\right)\right] \mathbb{P}\left(dx|\mathcal{F}_{-1}\right) \to 0. \tag{15}$$

For simplicity assume that $\mathfrak{G}$ only contains positive functions (otherwise deal with positive and negative part of each function separately). Therefore, for any finite M,

$$\begin{aligned}&\sup_{g\in\mathfrak{G}} \left| \int_{\mathcal{X}} g\left(x\right) \left[\mathbb{P}_n\left(dx|B_h\left(X_{-n}^{-1}\right)\right) - \mathbb{P}\left(dx|\mathcal{F}_{-1}\right)\right] \right| \\ \leq\ &\sup_{g\in\mathfrak{G}} \left| \int_{\mathcal{X}} \left(g\left(x\right) \wedge M\right) \left[\mathbb{P}_n\left(dx|B_h\left(X_{-n}^{-1}\right)\right) - \mathbb{P}\left(dx|\mathcal{F}_{-1}\right)\right] \right| \\ &+ \sup_{g\in\mathfrak{G}} \left| \int_{\mathcal{X}} \left[g\left(x\right) - \left(g\left(x\right) \wedge M\right)\right] \left[\mathbb{P}_n\left(dx|B_h\left(X_{-n}^{-1}\right)\right) - \mathbb{P}\left(dx|\mathcal{F}_{-1}\right)\right] \right| \\ =\ &\mathrm{I} + \mathrm{II}.\end{aligned}$$

Theorem 1 implies $\mathrm{I} \to 0$. Since $g \geq 0$,

$$\begin{aligned}
\sup_{g\in\mathfrak{G}} |g(x) - (g(x) \wedge M)| &= \sup_{g\in\mathfrak{G}} [g(x) - (g(x) \wedge M)] \\
&= \sup_{g\in\mathfrak{G}} [(g(x) - M) \{x \in \mathcal{X} : g(x) > M\}] \\
&\leq |(G(x) - M) \{x \in \mathcal{X} : G(x) > M\}| \\
&= \left(G(x) - G^M(x)\right).
\end{aligned}$$

Therefore, by the triangle inequality, Jensen inequality and then by the above display,

$$\begin{aligned}
\mathrm{II} &\leq \int_{\mathcal{X}} \sup_{g\in\mathfrak{G}} |g(x) - (g(x) \wedge M)| \, \mathbb{P}_n\left(dx|B_h\left(X_{-n}^{-1}\right)\right) + \int_{\mathcal{X}} \sup_{g\in\mathfrak{G}} |g(x) - (g(x) \wedge M)| \, \mathbb{P}(dx|\mathcal{F}_{-1}) \\
&\leq \left| \int_{\mathcal{X}} \left(G(x) - G^M(x)\right) \mathbb{P}_n\left(dx|B_h\left(X_{-n}^{-1}\right)\right) \right| + \left| \int_{\mathcal{X}} \left(G(x) - G^M(x)\right) \mathbb{P}(dx|\mathcal{F}_{-1}) \right|.
\end{aligned}$$

The first term can be made arbitrary small for M large enough, by asymptotic uniform integrability and similarly for the second term by (15). ■

Proof. [Corollary 2] Following the proof of Corollary 1 it is enough to show convergence for functions that are bounded and in $\mathfrak{G}$. Hence, by Lemma 10 in Sancetta (2007a) deduce that Theorem 1 implies the Corollary 2 (e.g. Sancetta, 2007b, section 3.3, for more details). ■

Proof. [Lemma 1]Note that for any real variables $(z_i)_{i\geq 1}$ and summable constants $(a_i)_{i\geq 1}$,

$$\left\{ \sum_{i\geq 1} z_i \leq \sum_{i\geq 1} a_i \right\} \supset \left\{ \bigcap_{i\geq 1} \{z_i \leq a_i\} \right\}.$$

To see this, consider

$$\left\{ \sum_{i\geq 1} z_i \leq \sum_{i\geq 1} a_i \right\} \supset \left\{ \{z_1 \leq a_1\} \cap \left\{ \sum_{i\geq 2} z_i \leq \sum_{i\geq 2} a_i \right\} \right\}$$

and proceed by induction. Hence deduce

$$\left\{ d\left(x_{-i}, X_{-(s+i)}\right) \leq \frac{h}{2i^2\lambda^i}; i = 1, \ldots, (n-s) \right\} \subset \left\{ \mathbf{d}_\lambda\left(x_{-(n-s)}^{-1}, X_{-n}^{-(1-s)}\right) \leq h \right\}$$

by letting $\sum_{i\geq 1} a_i = h > h\sum_{i\geq 1} i^{-2}/2$ and $z_i = \lambda^i d\left(x_{-i}, X_{-(s+i)}\right)$ in the first two displays. By Condition 5, with no loss of generality assume that $d \leq C = 1$ so that

$$\{y \in \mathcal{X} : d(x,y) \leq 1\} = \mathcal{X}.$$

Letting $I = I_{(h,\lambda)}$ be the smallest integer such that $h/\left(2I^2\lambda^I\right) \geq 1$, the previous display implies

$$\left\{d\left(x_{-i}, X_{-(s+i)}\right) \leq \frac{h}{2i^2\lambda^i}; i = 1, \ldots, (n-s)\right\} = \left\{d\left(x_{-i}, X_{-(s+i)}\right) \leq \frac{h}{2i^2\lambda^i}; i = 1, \ldots, I\right\}$$

and this last display implies

$$\sum_{s=1}^{n}\left\{\mathbf{d}_\lambda\left(x_{-(n-s)}^{-1}, X_{-n}^{-(1+s)}\right) \leq h_n\right\} \geq \sum_{s=1}^{n-I_{(h,\lambda)}}\left\{d\left(x_{-i}, X_{-(s+i)}\right) \leq \frac{h}{2i^2\lambda^i}; i = 1, \ldots, I\right\}. \tag{16}$$

From the definition of I, note that for any $\epsilon > 0$,

$$I = O\left(\left[\frac{\ln(1/h)}{\ln(1/\lambda)}\right]^{1/(1+\epsilon)}\right)$$

so that, for n large enough, $n - I > 0$. Define $Y_{-s}^I := \left(X_{-(s+1)}, X_{-(s+2)}, \ldots, X_{-(s+I)}\right)$. Then, the right hand side of (16) is the number of times (out of $n - I$ steps) the ergodic process $\left(Y_s^I\right)_{s\geq 0}$ visits open sets of positive radius, induced by d in each coordinate, and centered at x_{-I}^{-1}. By stationarity and ergodicity the process is recurrent and the number of visits of any open set centered at x_{-I}^{-1} goes to infinity as $n \to \infty$ for $\mathbb{P}$ almost all x_{-I}^{-1} by Poincare Recurrence Theorem (e.g. Theorem 6.4.1 in Gray 1998). Since h is arbitrary we can let $h = h_n \to 0$ slowly enough such that $\left(n - I_{(h,\lambda)}\right) \to \infty$ to deduce the final result. ■

References

[1] Backus, D. K., and A. W. Gregory (1992): "Theoretical Relations Between Risk Premiums and Conditional Variances," Working Paper EC–92–18, Stern School of

Business, NYU.

[2] Bandi, F. (2004). On persistence and nonparametric estimation (with an application to stock return predictability). Working paper.

[3] Clarkson, J.A. and C.R. Adams (1933) On Definitions of Bounded Variation for Functions of Two Variables. *Transactions of the American Mathematical Society* 35, 824-854.

[4] Collomb, G. (1985). Nonparametric time series analysis and prediction: uniform almost sure convergence. *Statistics* 2, 197-307.

[5] Dawid, A.P. (1997) Prequential analysis. In S. Kotz, C.B. Read and D.L. Banks (eds.), Encyclopedia of Statistical Sciences Volume 1, 464-470. Wiley.

[6] Devroye, L. (1981) On the almost everywhere convergence of nonparametric regression function estimates. *Annals of Statistics* 9, 1310-1319.

[7] Dudley, R.M. (2002) *Real analysis and probability.* Cambridge: Cambridge University Press.

[8] Ferraty, F., and P. Vieu (2007). *Nonparametric Functional Data Analysis.* Springer Verlag, Berlin.

[9] Gennotte, G., and T. Marsh (1988): "Valuations in Economic Uncertainty and Risk Premiums on Capital Assets," manuscript, UC Berkeley.

[10] Granger C.W.J. and Y. Jeon (2004) Thick Modeling. *Economic Modelling* 21, 323-343.

[11] Gray, R. (1998) *Probability, Random Processes, and Ergodic Properties.* New York: Springer. Revised version downloadable: <http://www-ee.stanford.edu/~gray/arp.pdf>.

[12] Gyorfi, L. , G. Morvai and S. Yakowitz (1998) Limits to consistent on-line forecasting for ergodic time series. *IEEE Transactions on Information Theory* 44, 886-892.

[13] Gyorfi, L., M. Kohler, A. Krzyzak and H. Walk (2002) *A Distribution Free Theory of Nonparametric Regression.* New York: Springer-Verlag.

[14] Hall, P. and C.C. Heyde (1980) *Martingale limit theory and its application.* New York: Academic Press.

[15] Joe, H. (1997) *Multivariate Models and Dependence Concepts.* London: Chapman and Hall.

[16] Karlsen, H.A. and D. Tjostheim (2001). Nonparametric estimation in null recurrent time series. *Annals of Statistics* 29, 372-416.

[17] Lenze, B (2003) On the Points of Regularity of Multivariate Functions of Bounded Variation. *Real Analysis Exchange* 29, 646-656.

[18] Linton, O. and E. Mammen, (2005), Estimating semiparametric ARCH(∞) models by kernel smoothing, *Econometrica*, **73**, 771-836.

[19] Linton, O.B. and J.P. Nielsen, (1995), A kernel method of estimating structured nonparametric regression using marginal integration, *Biometrika*, **82**, 93-100.

[20] Linton, O. B. and B. Perron (2003): The Shape of the Risk Premium: Evidence from a Semiparametric Generalized Autoregressive Conditional Model. *Journal of Business & Economic Statistics*, 354-367.

[21] Lu, Z., and O.B. Linton (2007) Asymptotic Distributions for Local Polynomial Nonparametric Regression Estimators under weak dependence *Econometric Theory* 23, 37-70.

[22] Mammen, E., O.B. Linton and J.P. Nielsen, (1999), The existence and asymptotic properties of a backfitting projection algorithm under weak conditions, *Annals of Statistics,* **27**, 1443-1490.

[23] Masry, E. (1996), "Multivariate local polynomial regression for time series: Uniform strong consistency and rates," *J. Time Ser. Anal.* 17, 571-599.

[24] Masry, E. and Fan, J. (1997). Local polynomial estimation of regression functions for mixing processes. *Scandinavian Journal of Statistics* **24**, 165-179

[25] Masry, E. (2005). Nonparametric regression for dependent functional data: asymptotic normality. *Stochastic Processes and their Applications* 115 (1), 155-177.

[26] Morvai, G., S. Yakowitz and L. Gyorfi (1996) Nonparametric inference for ergodic, stationary time series. *Annals of Statistics* 24, 370-379.

[27] Morvai, G., S. Yakowitz and P. Algoet (1997) Weakly convergent nonparametric forecasting of stationary time series. *IEEE Transation on Information Theory* 43, 483-498.

[28] Morvai, G. and B. Weiss (2005) Prediction for discrete time series. *Probability Theory and Related Fields* 132, 1-12.

[29] Pagan, A.R., and Y.S. Hong (1991): Nonparametric Estimation and the Risk Premium. In W. Barnett, J. Powell, and G.E. Tauchen (eds.) *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, Cambridge University Press.

[30] Pagan, A.R., and A. Ullah (1988): The econometric analysis of models with risk terms. *Journal of Applied Econometrics* 3, 87-105.

[31] Petrov, V. (1994) *Limit Theorems of Probability Theory.* Oxford: Oxford University Press.

[32] Phillips, P.C.B. and J.Y. Park (1998). Nonstationary density estimaiton and kernel autoregression. *Cowles Foundation Discussion Paper no.* 1181.

[33] Rio, E. (2000) *Théorie Asymptotique des Processus Aléatoires Faiblement Dépendants.* Paris: Springer.

[34] Robinson, P.M. (1983). Nonparametric estimation for time series models. *Journal of Time Series Analysis* 4, 185-208.

[35] Romano, J.P. and M. Wolf (2005) Stepwise multiple testing as formalized data snooping. *Econometrica* 73, 1237-1282.

[36] Sancetta (2007a) Weak Convergence of Laws on $\mathbb{R}^K$ with Common Marginals. *Journal of Theoretical Probability* 20, 371-380. Downloadable: <http://arxiv.org/abs/math.PR/0606462>.

[37] Sancetta (2007b) Nearest Neighbor Conditional Estimation for Harris Recurrent Markov Chains. Preprint. Downloadable: <http://www.sancetta.googlepages.com/academicpublications>.

[38] Schafer, D. (2002) Strongly consistent on-line forecasting of centered Gaussian processes. *IEEE Transactions on Information Theory* 48, 791-799.

[39] Seillier-Moiseiwitsch, F. and A.P. Dawid (1993) On testing the validity of sequential probability forecasts. Journal of American Statistical Association 88, 355-359.

[40] Stone, C.J. (1985) Additive regression and other nonparametric models. *Annals of Statistics* 13, 685-705.

[41] Tong, H. (1990). *Non-linear Time Series: A Dynamical System Approach*. Clarendon Press, Oxford.

[42] Van der Vaart, A. and J.A. Wellner (2000) *Weak Convergence and Empirical Processes*. Springer Series in Statistics. New York: Springer.